

**Hardware & Software for Big Data Mod (b)**

**M.s DATA SCIENCE**

**Prof. Flora Amato**

**June 2025**

**Final Project**

**Smart Water Management - Lakes**

**Project members:**

**Zaighum Jawad Aslam**

**Samra Amjed**

**Airaf Siddique**

Table of Contents

Contents

[1. Dataset Overview 4](#_Toc200954407)

[1.1 Real Problem Presentation 4](#_Toc200954408)

[1.2 Problem Statement & Objective 4](#_Toc200954409)

[1.3 The Dataset and its understanding 4](#_Toc200954410)

[Dataset Variables: Detailed Definition 5](#_Toc200954411)

[Column Name 5](#_Toc200954412)

[Description 5](#_Toc200954413)

[Data Type 5](#_Toc200954414)

[1.4 Summary Statistics of Dataset 6](#_Toc200954415)

[2. Exploratory Data Analysis (Data Loading, Descriptive Statistics & Data Analysis) 6](#_Toc200954416)

[2.2 Histograms: 7](#_Toc200954417)

[2.3 Anomaly Detection Using Z-Score Method 9](#_Toc200954418)

[2.4 Correlation Heatmap 10](#_Toc200954419)

[3. Stationarity Check using Augmented Dickey-Fuller Test 12](#_Toc200954420)

[4. Modeling 12](#_Toc200954421)

[4.1 ARIMA Modeling for Lake Level Forecasting 12](#_Toc200954422)

[4.1.1 Model Evaluation 13](#_Toc200954423)

[5. Forecasting Results 13](#_Toc200954426)

[6. Model Performance 21](#_Toc200954427)

[7. Conclusion 22](#_Toc200954428)

[8. Future Work 23](#_Toc200954429)

# Dataset Overview

## Real Problem Presentation

**Introduction**

Water is one of the most critical natural resources, and its sustainable management is essential for environmental balance and human survival. In the context of smart water management, accurate forecasting of water levels and flow rates in natural water bodies plays a vital role in ensuring optimal water usage, infrastructure planning, and resource conservation. As each waterbody—such as lakes, rivers, aquifers, and springs—exhibits unique hydrological behaviors, their predictive analysis must consider these inherent differences.

This project focuses on time-series forecasting for one such waterbody: a lake. The analysis is centered on understanding and predicting key attributes such as **lake water levels** and **flow rates** using historical data. The dataset used originates from the *Acea Smart Water Analytics Challenge*, hosted on Kaggle, and has been curated to reflect daily observations of environmental features like rainfall, temperature, flow rate, and lake level. These observations span multiple years and offer valuable insights into seasonal and long-term trends affecting water availability.

The primary aim of this analysis is to utilize advanced forecasting models such as **ARIMA** and **Prophet** to project future lake levels and flow rates. This serves practical decision-making scenarios for water management authorities, especially during critical periods like summer droughts or flood-prone monsoon months. Additionally, the project undertakes comprehensive exploratory data analysis (EDA), anomaly detection, and statistical validation (e.g., stationarity tests) to ensure that the time-series models are applied correctly and reliably.

Through this investigation, the project contributes to developing intelligent systems that can aid policy-makers and water engineers in understanding hydrological trends and preparing for future resource demands.

## Problem Statement & Objective

The objective of this project is to develop a robust predictive model to forecast lake water levels based on past hydro-meteorological parameters. Key goals include:

* Understand the relationship between rainfall, temperature, and lake inflow with water level
* Identify seasonal patterns and trends
* Build a predictive model using ARIMA and SARIMAX
* Evaluate model accuracy using metrics like RMSE, MAE, MAPE, and R-square

## The Dataset and its understanding

The dataset is a time series of daily water and climate measurements. It was originally taken from a smart water monitoring system and consists of 6,602 observations recorded on a daily frequency. It covers several hydrological and environmental indicators necessary to forecast lake water levels.

Each record corresponds to one day, capturing the following key measurements.

The project commenced with a thorough data understanding phase, utilizing the dataset lakes\_fe.csv, which contains measurements of various environmental features related to a lake water body. The dataset was loaded using the pandas library and previewed to understand its structure. Initial inspection showed relevant columns such as Mean\_Rainfall, Mean\_Temp, Actual\_Flow\_Rate, and Actual\_Lake\_Level, all of which appeared numerically consistent and vital for downstream modeling.

To assess data quality, missing value analysis was conducted, and the results confirmed that the dataset had minimal or no missing values, making it well-suited for modeling without extensive preprocessing. Summary statistics such as mean, standard deviation, minimum, and maximum values were computed to identify potential anomalies, skewness, and outliers in each variable.

Following this, the data types of all columns were inspected. Numerical features were confirmed to be of type float64, which is optimal for most statistical and machine learning algorithms. The Date column was excluded from float-based operations to maintain its role as a temporal index.

To visually identify outliers, **boxplots** were created for all float-type features. These boxplots helped in highlighting the spread and possible extreme values within each variable, giving an early indication of data distribution and potential anomalies.

Additionally, **histograms** were plotted for each float column to visualize the frequency distribution. For Mean\_Rainfall, Mean\_Temp, and Actual\_Flow\_Rate, histograms were constructed using bins of size 10 over a typical range of 0–100, offering insights into the general spread of values. A more fine-grained histogram was created for Actual\_Lake\_Level, using bin sizes of 1 across a narrowed range (240–255), revealing subtle shifts and patterns in lake level observations.

These exploratory visualizations not only validated the quality and usability of the dataset but also helped in establishing initial intuitions about the nature of each variable—crucial before applying any predictive modeling techniques.

|  |
| --- |
| Data types of the columns:  Date int64  Mean\_Rainfall float64  Mean\_Temp float64  Actual\_Flow\_Rate float64  Actual\_Lake\_Level float64  dtype: object |

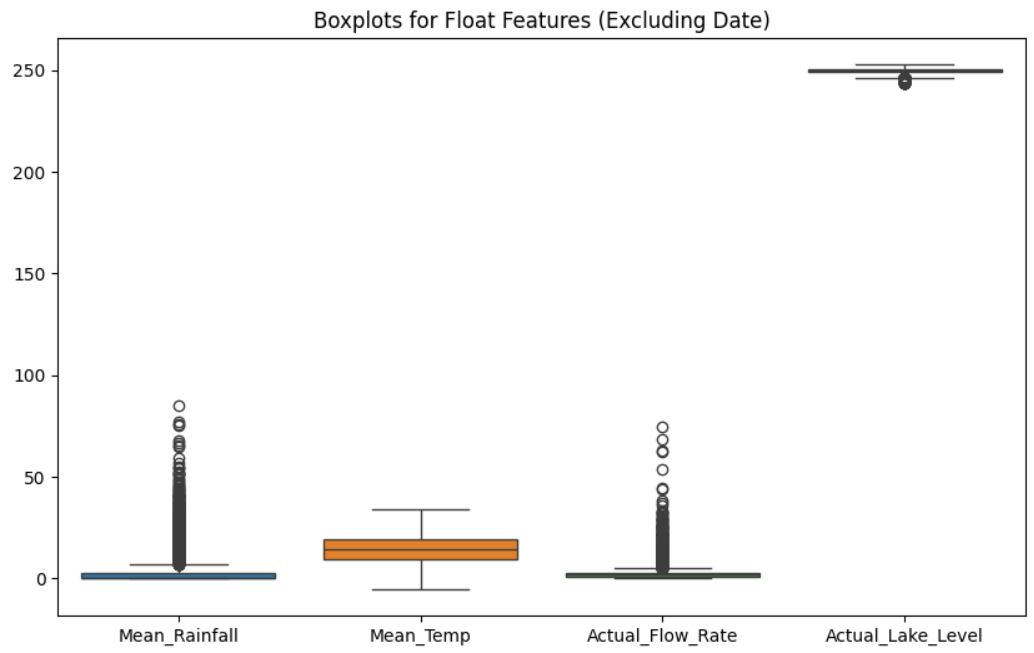


Figure 1: Box Plot

## Dataset Variables: Detailed Definition

Below variables provide a multi-dimensional view of the water system behavior over time.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| Date | Timestamp of the measurement (daily frequency) | datetime |
| Mean\_Rainfall | Average rainfall measured in mm for the day | float |
| Mean\_Temp | Mean ambient temperature recorded on the day (°C) | float |
| Actual\_Flow\_Rate | Volume of water flowing into the lake (cubic meters per second approx.) | float |
| Actual\_Lake\_Level | Observed lake level in meters from a reference point | float |

## Summary Statistics of Dataset

* **Date**: Ranges from 0 to 6602, with a mean of 3301, representing a time index.
* **Mean\_Rainfall**: Ranges from 0 to 85.12, with a mean of 2.86.
* **Mean\_Temp**: Ranges from -5.35 to 34°C, with a mean of 14.53°C.
* **Actual\_Flow\_Rate**: Ranges from 0.45 to 74.65, with a mean of 2.75.
* **Actual\_Lake\_Level**: Ranges from 243.53 to 252.76, with a mean of 249.56.

# Exploratory Data Analysis (Data Loading, Descriptive Statistics & Data Analysis)

The dataset consists of daily hydrological and meteorological records, which are essential for analysing and forecasting lake water levels. The initial dataset was loaded using pandas and examined to understand its structure, completeness, and data types.

Upon inspection:

* The dataset contained 6,602 records and 5 columns.
* The 'Date' column was converted to datetime format and set as the index for time series operations.
* Other columns like rainfall, temperature, flow rate, and lake level were of numeric (float64) types.

To ensure data integrity, a thorough check for missing values was performed. The dataset was already clean with no significant null values. Basic statistics (mean, median, standard deviation) were generated using describe (), which helped in identifying data scale and variance across features.

**2.1 Outlier Detection & their removal:**

Outlier detection was conducted using the IQR (Interquartile Range) method. Each numeric column (Rainfall, Temperature, Flow Rate, and Lake Level) was examined for extreme values beyond the typical 1.5×IQR range. These outliers were then removed to enhance model robustness and prevent skewed predictions. To ensure the quality of data before applying any predictive modeling, we conducted anomaly detection using the Z-score method. Z-scores were computed for the Actual\_Lake\_Level and Actual\_Flow\_Rate variables to identify values that significantly deviate from the mean. A threshold of ±3 standard deviations was used to flag anomalous data points, which is a standard statistical approach for outlier detection. These anomalies were then visually highlighted using scatter plots overlaid on time series graphs. The top subplot displayed Actual\_Lake\_Level along with its anomalies in red, while the bottom subplot did the same for Actual\_Flow\_Rate. This analysis helped in identifying extreme fluctuations that may impact the stability of the forecasting models, allowing us to consider whether these anomalies should be retained or corrected based on domain relevance.

To complement the Z-score anomaly analysis, we also performed outlier detection using the Interquartile Range (IQR) method. For each numerical feature in the dataset (excluding the date), we calculated the first (Q1) and third quartiles (Q3) to determine the IQR. Any data point falling below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR was flagged as a potential outlier. This robust statistical technique allowed us to summarize and quantify the extent of deviation in each column. The results include the quartile boundaries, IQR, and outlier counts for all float-type variables. This procedure aids in understanding data spread and assessing whether these extreme values should be addressed before applying predictive modeling techniques.

After outlier detection and their removal the original records of data which were 6602 remained 4697.

## 2.2 Histograms:

Histograms for each feature were generated with custom bin sizes to better observe distribution shape. Key takeaways from the plots:

* Actual\_Flow\_Rate had outliers but also maintained a relatively long tail.
* Rainfall had the most right-skewed shape with many zero or near-zero entries.
* Lake Level displayed a bell-shaped distribution, validating its continuous and relatively stable nature.

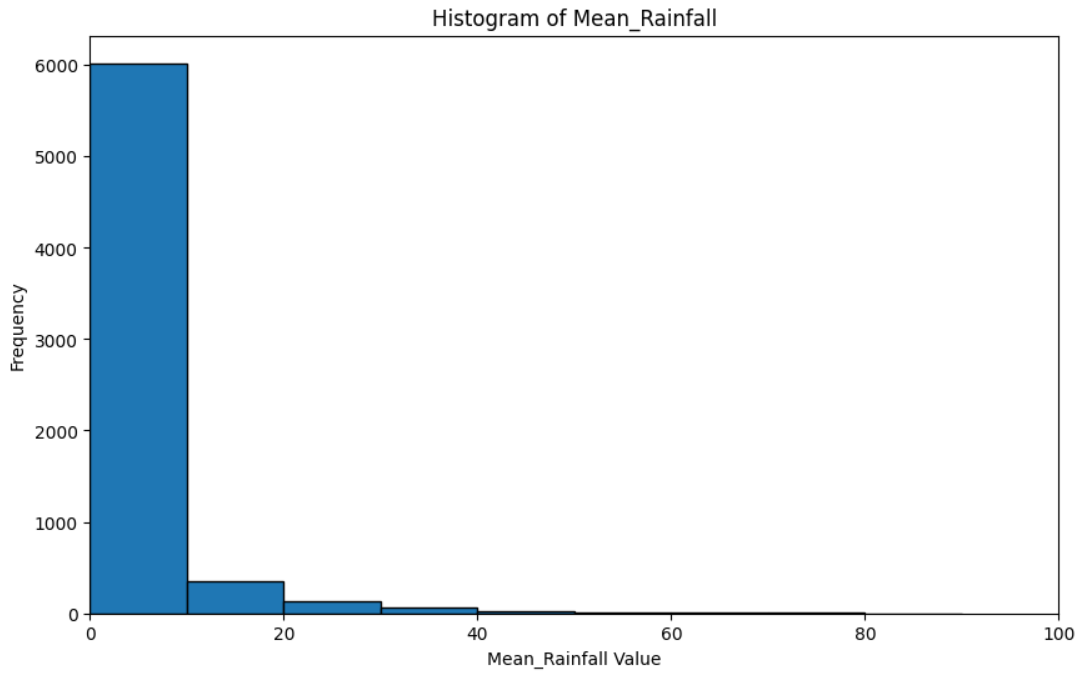


Figure 2: Histogram Of Rainfall

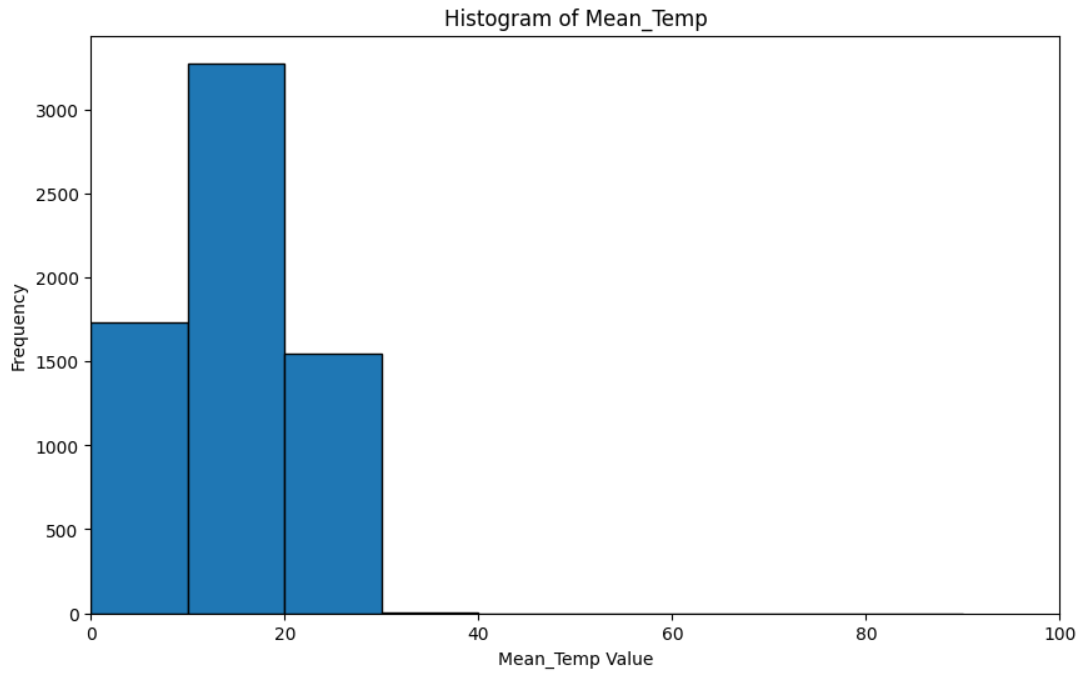


Figure 3: Histogram Of Temperature

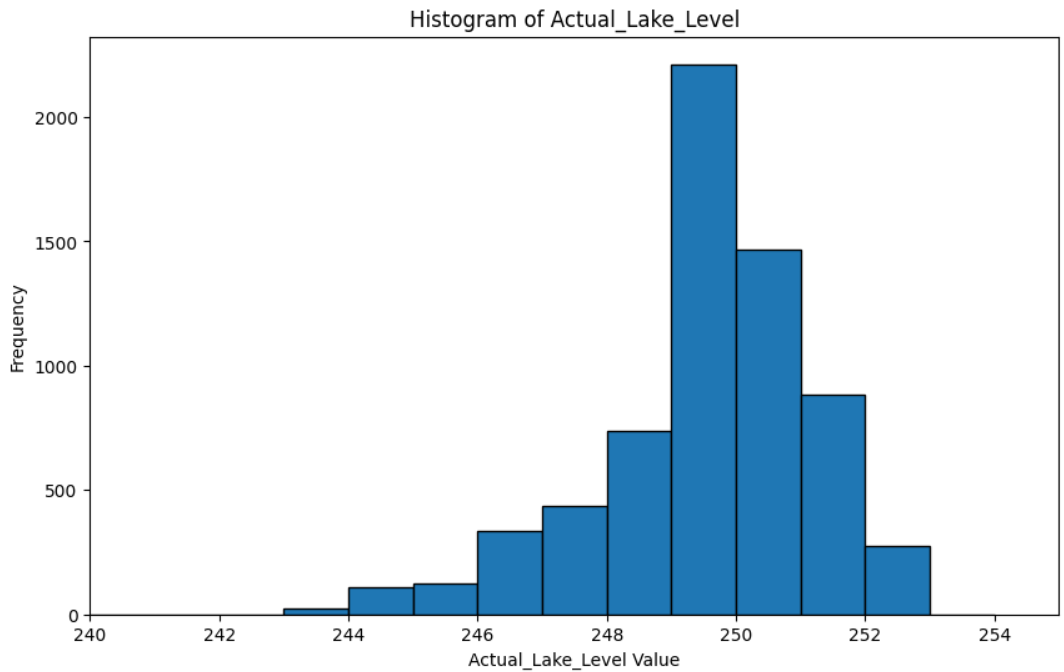


Figure 4:Histogram Of Lake Level

## 2.3 Anomaly Detection Using Z-Score Method

To further strengthen data quality and understand rare events, we implemented anomaly detection using the Z-score method. The Z-score standardizes values by subtracting the mean and dividing by the standard deviation, allowing us to detect values that deviate significantly from the norm. We applied this technique specifically to the ‘Actual\_Lake\_Level’ and ‘Actual\_Flow\_Rate’ variables, both of which are critical to water management. A Z-score threshold of ±3 was used, meaning any data point that deviated by more than 3 standard deviations from the mean was flagged as an anomaly. This technique highlighted extreme peaks or dips in lake levels and inflows that could be attributed to abnormal rainfall events, infrastructure issues, or sensor malfunctions. The detected anomalies were visualized using time series plots, where anomalous points were highlighted in red. These visuals helped contextualize when and how often such events occurred within the dataset. This step not only improved understanding of the data but also informed decisions about whether such outliers should be retained, flagged, or removed during modelling.

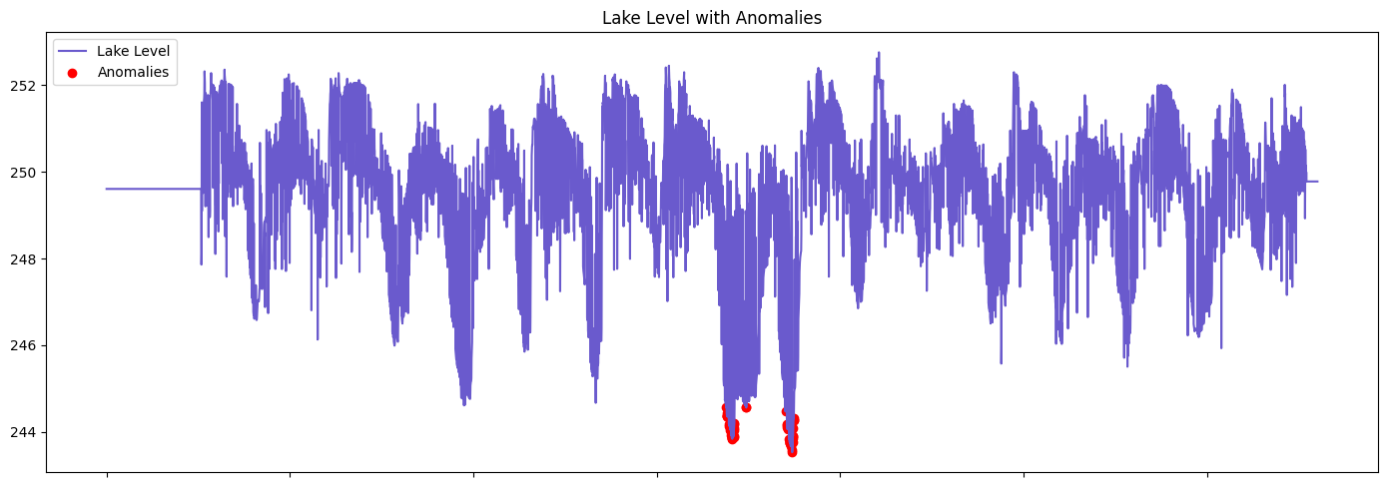


Figure 5:Anamolies Representation (Lake Level)

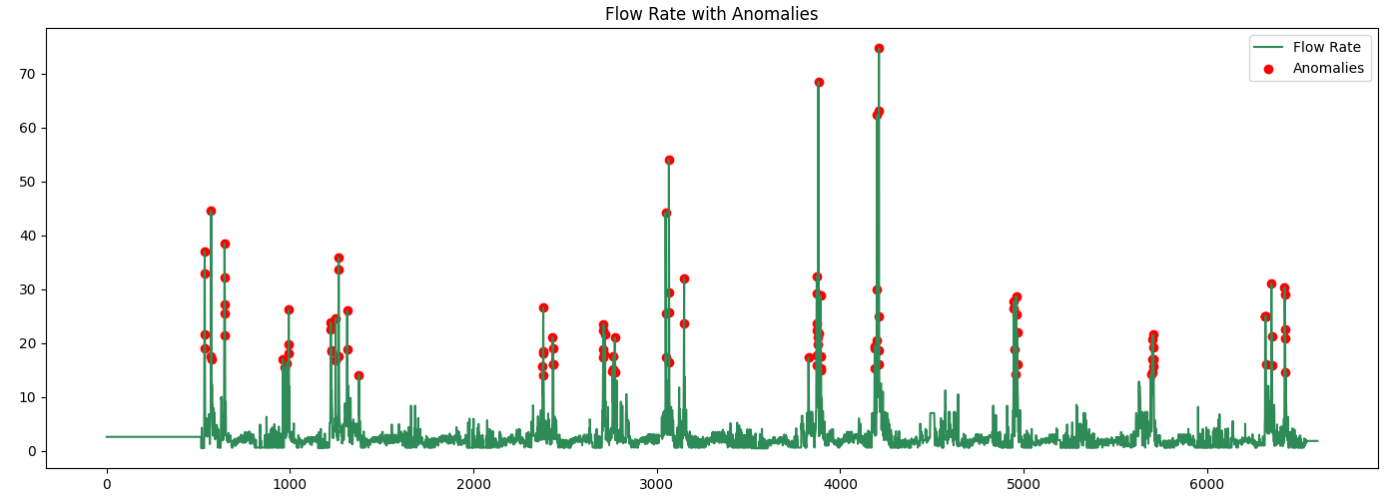


Figure 6: Anamolies Representation (Flow Rate)

## 2.4 Correlation Heatmap

To explore inter-variable relationships, a correlation heatmap was generated using Pearson correlation coefficients. This matrix provides insights into the linear associations between the main numerical features: Mean\_Rainfall, Mean\_Temp, Actual\_Flow\_Rate, and Actual\_Lake\_Level.

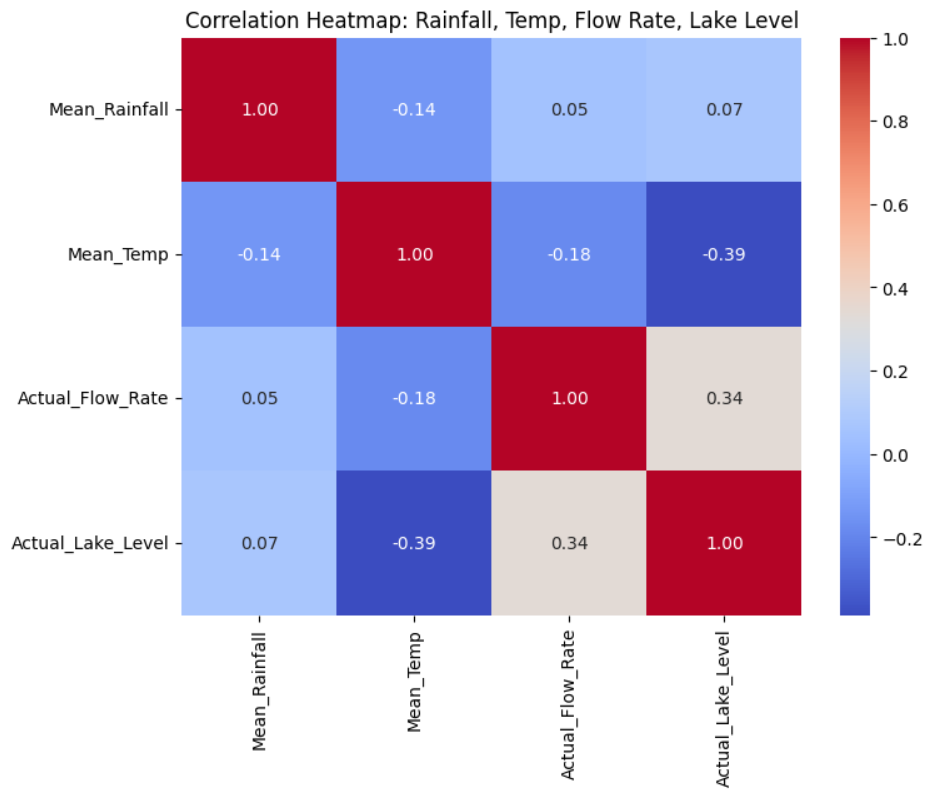


Figure 7: Correlation Heatmap Representation

To investigate the interrelationships among the core numerical variables in the dataset, we generated a correlation heatmap focusing on four key features: Mean\_Rainfall, Mean\_Temp, Actual\_Flow\_Rate, and Actual\_Lake\_Level. The Pearson correlation coefficient was used to quantify the linear dependency between variables. The results reveal a mild positive correlation (0.34) between Actual\_Flow\_Rate and Actual\_Lake\_Level, suggesting that higher flow rates are generally associated with higher lake levels. On the other hand, Mean\_Temp exhibits a moderate negative correlation (-0.39) with lake levels, indicating that increased temperatures may be linked to reduced lake levels, possibly due to increased evaporation or seasonal effects. Interestingly, Mean\_Rainfall shows only weak correlations with other features, particularly a near-zero relationship with lake level (0.07), which might be attributed to delayed or indirect rainfall effects. This analysis helps in identifying relevant predictors and understanding potential redundancies, which is crucial for downstream modeling and forecasting.

This correlation analysis helped identify which features may serve as effective predictors in the SARIMAX modeling phase and informed the decision to include Flow Rate and Temperature as key exogenous variables.

Following this, a new CSV file containing the cleaned dataset ("cleaned\_lakes\_data\_2012\_daily.csv") was saved and used for further analysis and modelling.

This comprehensive pre-processing ensured the dataset was consistent, statistically sound, and ready for exploratory analysis and time series modelling.

# Stationarity Check using Augmented Dickey-Fuller Test

To assess the stationarity of the Actual\_Lake\_Level time series—a critical assumption for time-series modeling we employed the Augmented Dickey-Fuller (ADF) test. This statistical test evaluates whether a unit root is present in the series, which would indicate non-stationarity. The ADF statistic obtained was **-6.0386**, and the associated p-value was **1.36e-07**, which is significantly below the conventional 0.05 threshold. Therefore, we reject the null hypothesis (H₀: the series has a unit root), concluding that the lake level series is stationary. This result supports the application of time-series models like ARIMA that assume the underlying series has stable mean and variance over time.

# Modeling

## 4.1 ARIMA Modeling for Lake Level Forecasting

After confirming the stationarity of the Actual\_Lake\_Level series, we proceeded with time series modeling using the ARIMA (AutoRegressive Integrated Moving Average) framework. Specifically, we implemented an **ARIMA(1, 0, 1)** model, where:

* **p = 1**: One autoregressive term,
* **d = 0**: No differencing (as the series was stationary),
* **q = 1**: One moving average term.

The model was trained on the full cleaned dataset after setting the Date column as the datetime index and sorting the data chronologically. To assess the model’s performance, we generated **in-sample predictions** over the entire dataset.

To begin the time-series forecasting process, we applied the ARIMA model to predict daily lake levels. After loading the cleaned dataset (cleaned\_lakes\_data\_2012\_daily.csv), we converted the Date column to datetime format and set it as the index to facilitate time-based analysis. The ARIMA model was initialized with the parameters *(1, 0, 1)*, representing the order of the autoregressive term, degree of differencing, and moving average term, respectively. This configuration was chosen after confirming stationarity using the Augmented Dickey-Fuller test. Once the model was fitted to the Actual\_Lake\_Level series, in-sample predictions were generated and plotted against the original data. The graph clearly demonstrates how closely the ARIMA model tracks the actual lake levels over time. To improve interpretability, the x-axis was formatted to display years only. This visual comparison highlights the ARIMA model’s capacity to capture seasonal and trend patterns in the lake level data with reasonable accuracy.

## Model Evaluation

# 

Figure 8: Fitting Of Arima On Lake Level

# Following the fitting of the ARIMA model, a future forecast was performed to estimate lake levels for the next 365 days beyond the available dataset. The model utilized the get\_forecast() method to project the expected lake levels along with the 95% confidence intervals for each day. A new date range was generated starting from the last available date in the dataset, ensuring that the forecast aligns with the temporal continuity of the original series. The resulting visualization overlays the historical lake levels with the model's forecast, visually distinguishing the predicted trend and the uncertainty bounds. The green shaded area in the graph represents the confidence interval, indicating the range within which future lake levels are likely to fall. This forecasting step is critical for water resource planning, as it provides an empirical basis for anticipating changes in lake volume and proactively managing infrastructure and policy.

# Forecasting Results

**Lake Levels**

**Arima:**

To forecast future lake levels, the ARIMA model trained on historical daily lake level data from 2012 onward was extended to predict the next 365 days. After fitting the ARIMA(1, 0, 1) model, we generated a forecast beginning the day after the last available observation in the dataset. The forecast output includes both the predicted mean values and a 95% confidence interval, allowing us to visualize potential uncertainty in future predictions.The visualization clearly distinguishes between actual historical data and future forecasted values. The green shaded area in the forecast graph represents the confidence interval, indicating the possible range within which the true lake level might fall. Notably, the model anticipates a relatively stable lake level with mild fluctuations over the forecast period. However, the widening of the confidence band over time highlights increasing uncertainty, a typical behavior in time-series forecasting. This result supports informed water management planning by offering a reliable short-term projection of water body behavior.

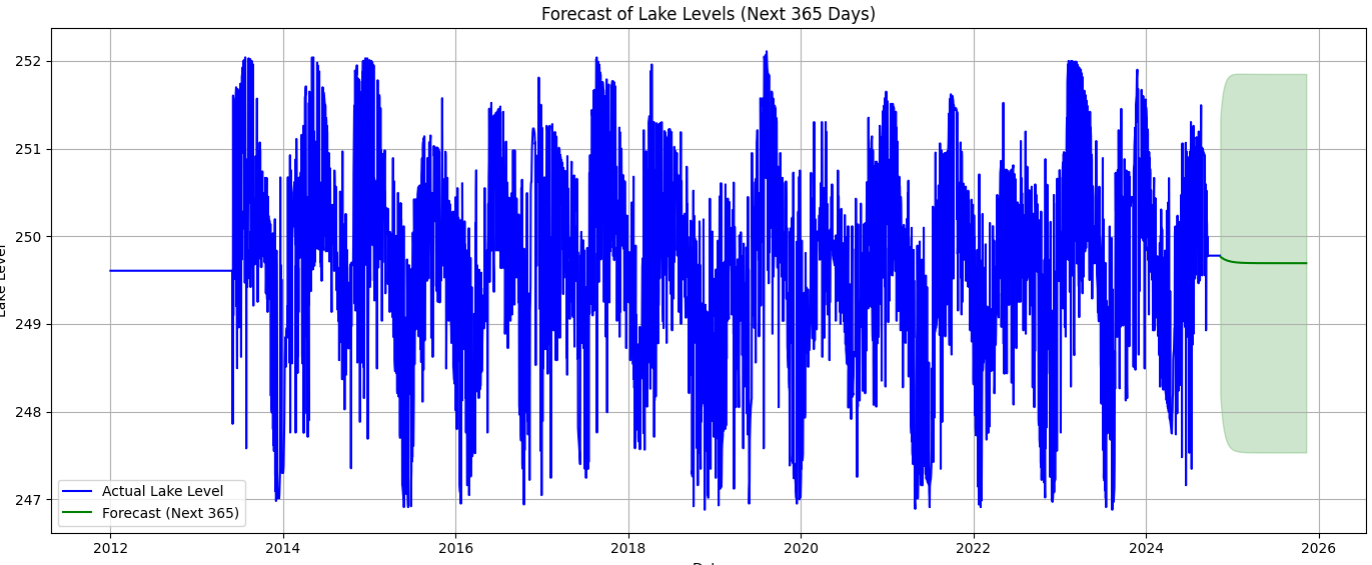


Figure 9: Forecasting Of Lake Levels

**Auto Correlation Function:**

To gain deeper insights into the temporal dependencies of lake level measurements, both Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were analyzed. The ACF plot illustrates the correlation of the time series with its own lagged values over a range of 40 lags. As shown in the ACF graph, there is a strong positive autocorrelation at lag 1, and the correlation gradually decays as the lag increases. This behavior suggests a clear persistence or memory in the series, where past values continue to influence future values over time. On the other hand, the PACF plot reveals the direct relationship between the series and its lagged values, controlling for the influence of intermediate lags. In the PACF graph, we observe significant spikes at lags 1 and 2, beyond which the correlations diminish sharply. This pattern typically indicates that an autoregressive model of order 1 or 2 (AR(1) or AR(2)) may be suitable for modeling the series. Together, these plots guide the selection of appropriate parameters for ARIMA modeling by helping to identify the number of autoregressive and moving average terms required for capturing the underlying data structure effectively.

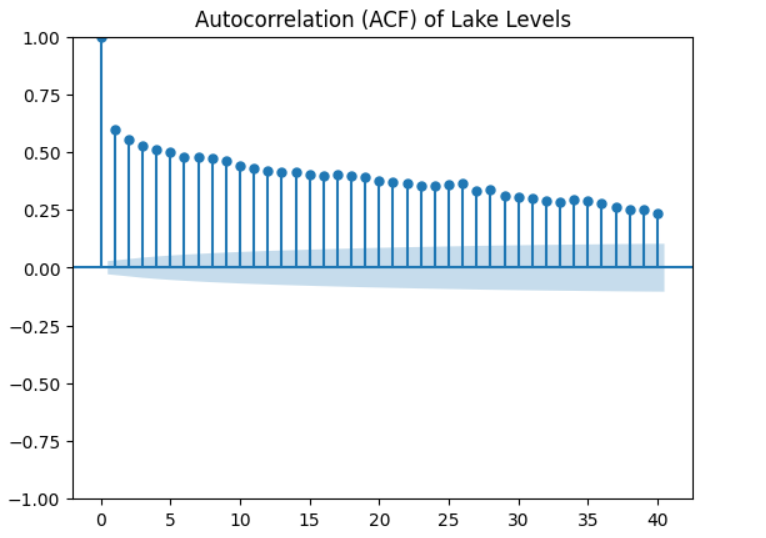


Figure 10: ACF Graph

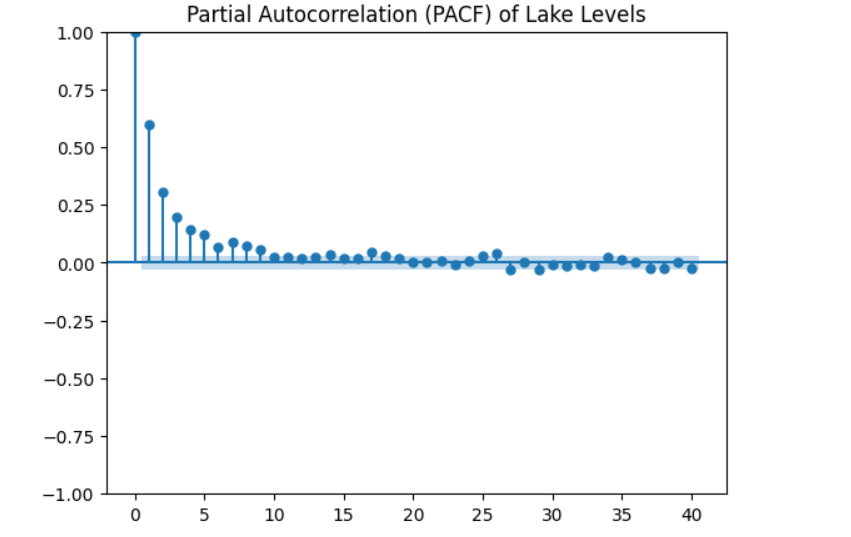


Figure 11: PACF Graph

**Prophet Model:**

To complement the ARIMA-based analysis and provide a robust time-series forecasting perspective, Facebook’s Prophet model was applied to predict lake levels. The dataset, containing daily observations from 2012 onwards, was restructured to fit Prophet’s required format with ds representing the date and y the lake level. Prophet, being well-suited for time series exhibiting strong seasonal effects and historical trends, was configured with yearly seasonality enabled. The model was trained on the historical data and used to forecast lake levels for the next 365 days. The resulting forecast plot provides both the predicted trend and a confidence interval visualized as a shaded region, offering interpretability and uncertainty estimation. The x-axis was formatted using yearly tick locators to enhance temporal readability across the forecast horizon (2012–2026). Prophet’s ability to decompose trends and handle time-series anomalies without extensive parameter tuning makes it a strong supplementary model alongside traditional statistical methods like ARIMA.

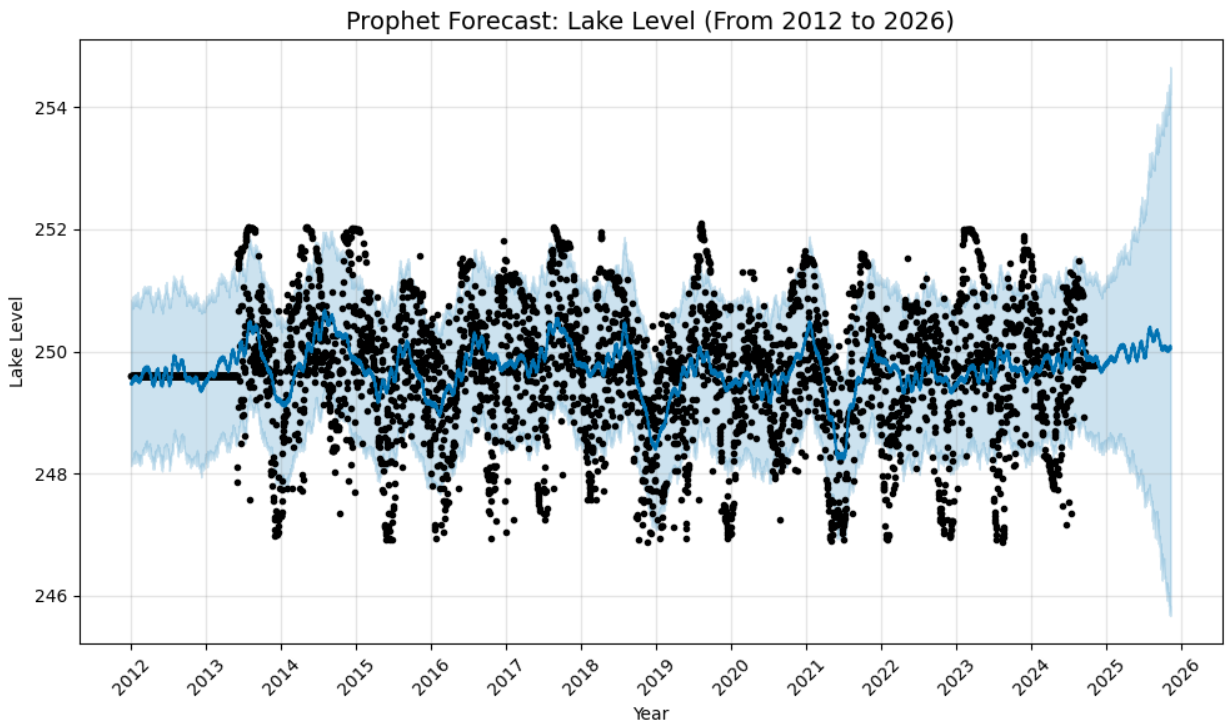


Figure 12: Prophets Prediction Of Lake Level

To better understand the underlying structure of the lake level data, Prophet’s decomposition functionality was applied. This process disaggregates the time series into three distinct components: trend, weekly seasonality, and yearly seasonality. The top subplot in the decomposition illustrates the long-term trend, which remains relatively stable with a slight upward movement toward the end of the forecast window. The weekly component, displayed in the second subplot, captures short-term variations throughout the week and shows that lake levels slightly increase towards the weekend, peaking on Saturdays. This might suggest operational or environmental changes over the week. The yearly component reveals strong periodic fluctuations, indicating the presence of seasonal effects likely caused by rainfall or climate patterns across the year. For better interpretability, these seasonalities were also plotted on the same scale as the original lake levels. By adding the trend back to the seasonal components, we visualized the actual expected magnitude of variation over time. This multi-scale decomposition enables us to identify the cyclical behaviors and isolate the long-term growth trend, making Prophet particularly useful for water body level prediction in hydrological applications.

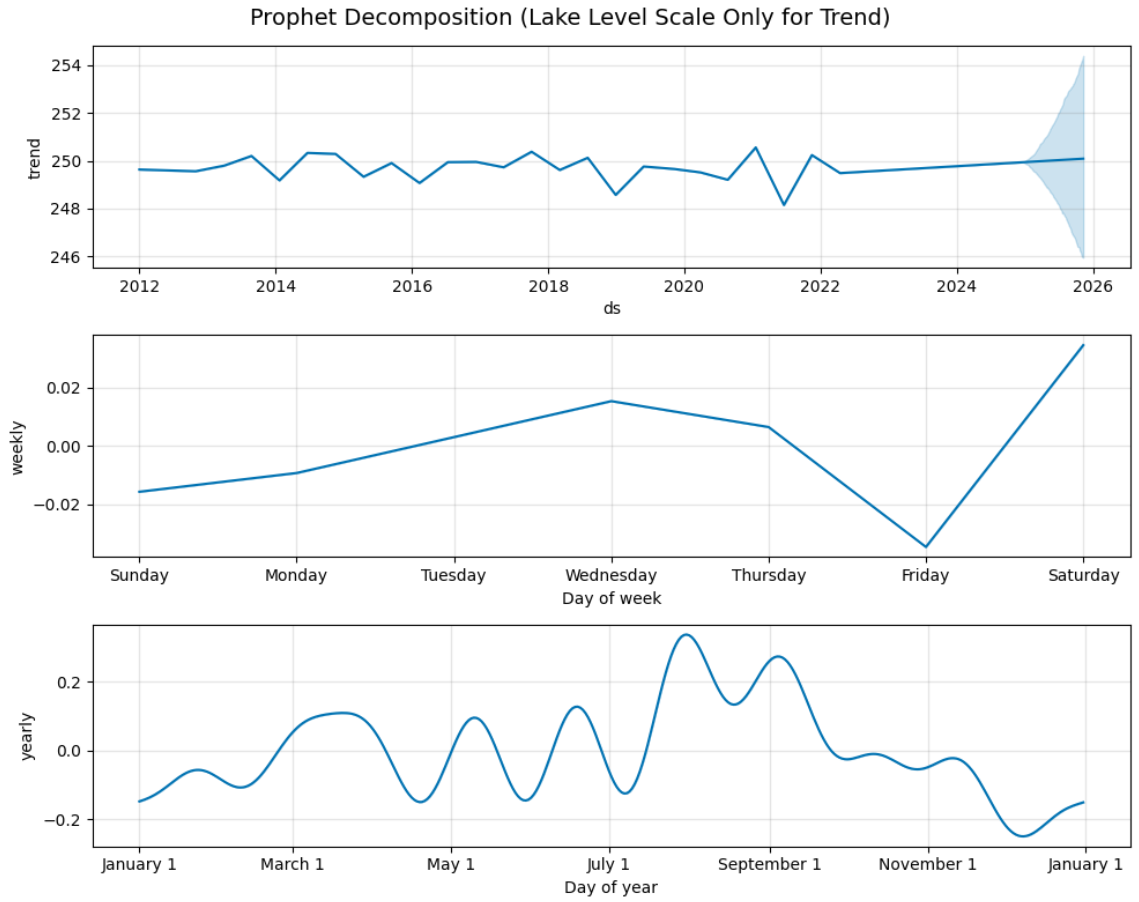


Figure 13: Diagnostics Of Prophet Model For Lake Level

**Flow Rate**

**Arima Model:**

In this section, the ARIMA model was employed to forecast the lake's flow rate for a future horizon of five years (2023 to 2027), using actual cleaned daily data starting from 2012. The time series was first prepared by assigning a proper datetime index, enabling ARIMA to recognize temporal patterns accurately. The ARIMA(1,0,1) configuration was selected for modeling based on previous exploratory analysis. After fitting the model to the historical Actual\_Flow\_Rate data, a forward prediction for 1825 days was generated. The forecast included both point predictions and a confidence interval to represent uncertainty in the future estimates. As shown in the resulting plot, the model is able to extend the observed temporal dynamics while capturing general trends. The forecast visually aligns with the long-term behavior of the flow rate, offering a reasonable estimate of expected values under current conditions. This forecast can assist in anticipating water availability and aid in planning for reservoir and infrastructure management.

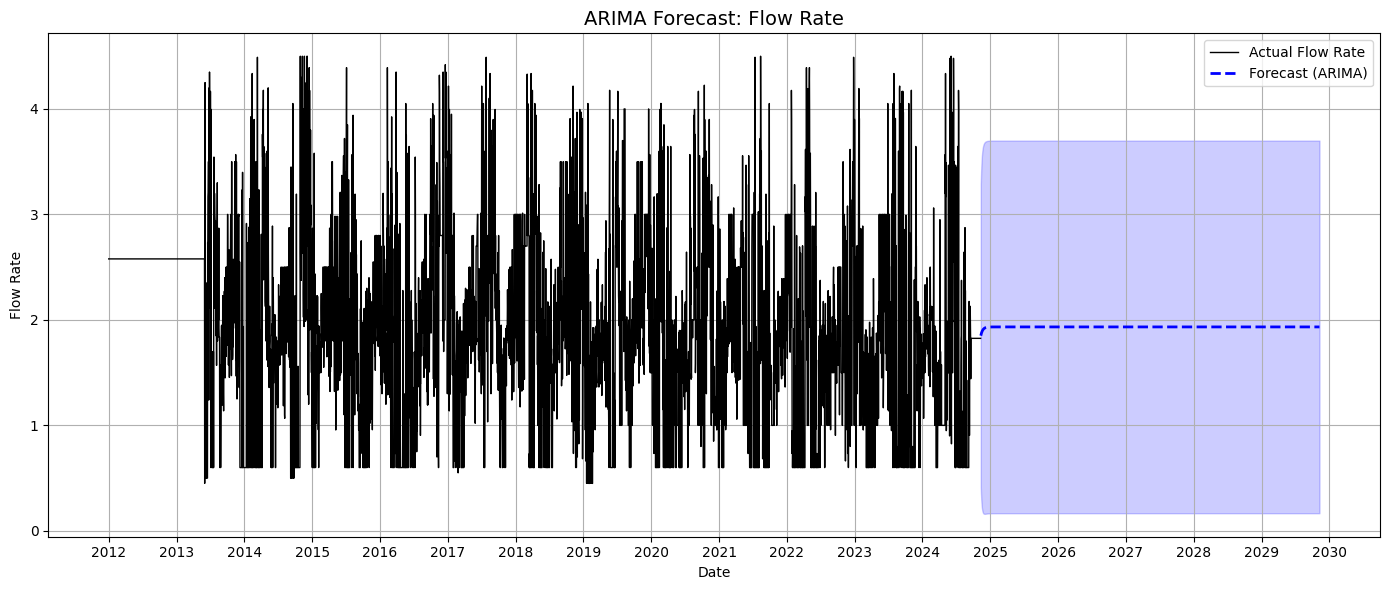


Figure 14: Arima Model (Flow Rate)

To complement ARIMA-based forecasting, the Facebook Prophet model was also employed for long-term flow rate prediction. Prophet is a robust time series forecasting tool designed to capture both trend and seasonality patterns. In this approach, cleaned daily flow rate data from 2012 onward was prepared for modeling. A new datetime index starting from January 1, 2012, was assigned to ensure consistent time intervals. The Prophet model was configured with yearly seasonality to reflect expected annual fluctuations in flow behavior. After training the model on the full dataset, a future forecast was generated for a five-year period (1825 days), covering 2023 to 2027. The resulting plot showed predicted flow rates along with uncertainty intervals, illustrating both central trends and variability in expected values. Additionally, Prophet’s decomposition plots revealed underlying components such as overall trend, weekly, and yearly seasonal influences, offering transparent insights into how the model interprets historical dynamics. This modeling effort aids in understanding and anticipating future changes in water flow, which is critical for environmental management and operational planning.

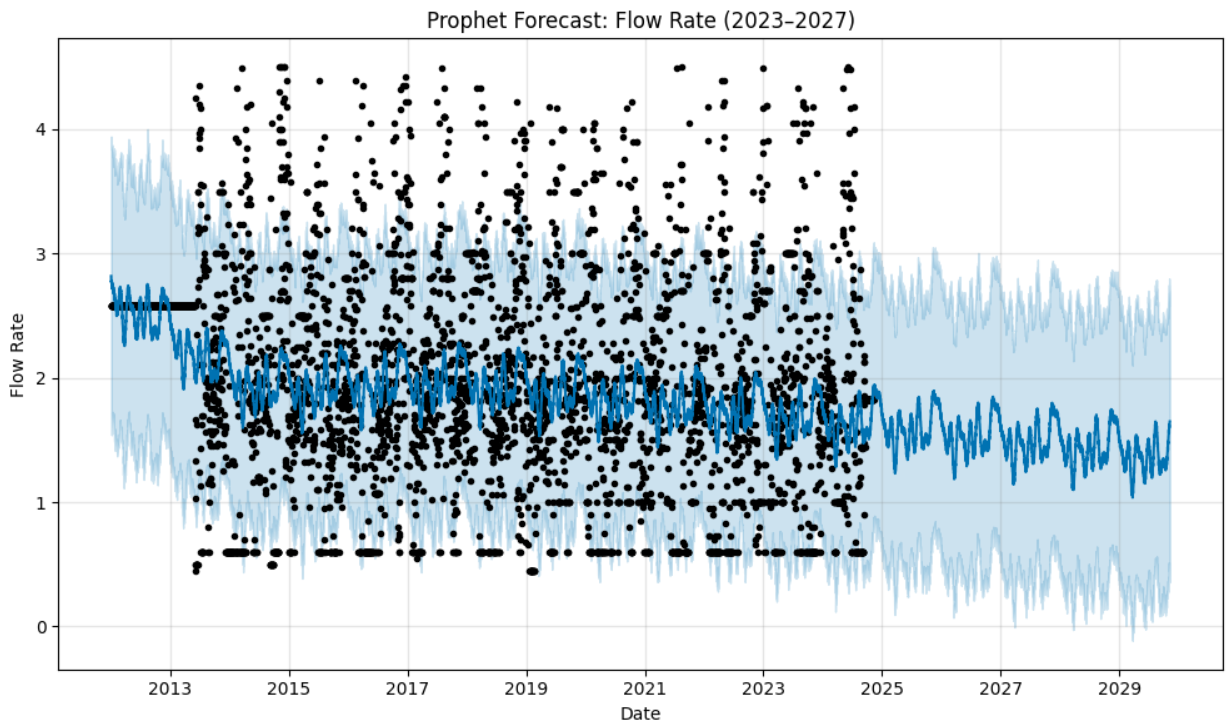


Figure 15: Prophet Model On Flow Rate Prediction

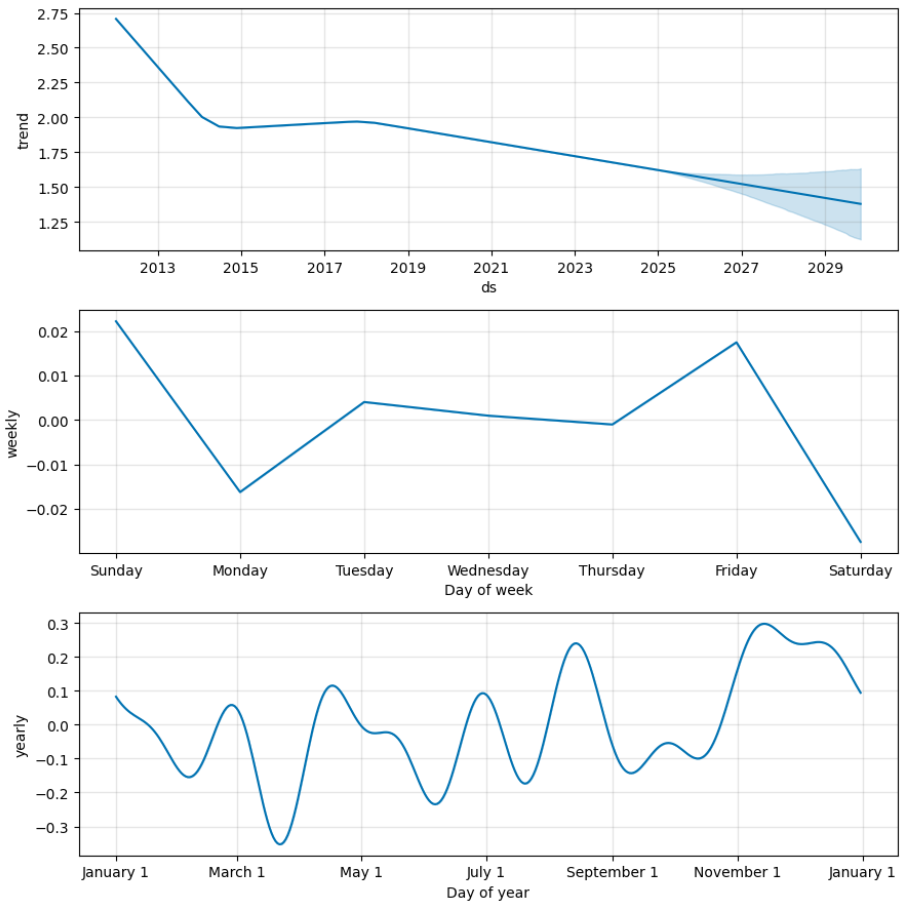


Figure 16: Prophet Model Diagnostics For Flow Rate

To further analyze the behavior of water flow rates over time, we employed the Prophet model to decompose the time series into its primary components: trend, weekly seasonality, and yearly seasonality. The top panel of the decomposition plot clearly illustrates a gradual declining trend in flow rate from 2012 to the projected future in 2029. This declining trend may indicate a long-term reduction in inflow or climatic changes affecting the water body. The middle panel represents the weekly seasonality component, revealing that flow rates tend to slightly increase toward the end of the week, peaking on Sundays and Fridays, and decreasing sharply on Mondays and Saturdays. This may correlate with human water usage patterns or natural fluctuations. The bottom panel shows the yearly seasonality, where multiple peaks and troughs are observed across the months. Notably, flow rates tend to increase during late summer and early winter, suggesting the influence of seasonal precipitation or meltwater patterns. This detailed decomposition not only improves interpretability but also supports strategic water management decisions by uncovering underlying temporal patterns in the flow rate data.

# Model Performance

**Arima Model:**

To assess the predictive accuracy of the ARIMA model on lake level forecasting, we evaluated the model's 365-day forecast using three standard performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²). The comparison was made between the model’s forecasted values and the actual lake level values over the last 365 days of available data. RMSE quantifies the average magnitude of the forecasting errors, with a lower value indicating better performance. MAE complements this by measuring the average absolute difference between predicted and actual values, offering a more interpretable error magnitude. R² evaluates the proportion of variance in the actual data explained by the model; a value closer to 1 indicates strong predictive capability. These metrics collectively help quantify how well the model can generalize and serve as an early validation of its forecasting reliability before future values are realized.

|  |
| --- |
| ARIMA Forecast Evaluation (Next 365 Days):  RMSE: 1.0060  MAE : 0.8000  R²: -0.0082 |

To assess the performance of the ARIMA model in forecasting lake levels, we evaluated its predictions over a 365-day horizon using standard regression metrics. The Root Mean Squared Error (RMSE) was found to be **1.0060**, indicating a small average deviation between the predicted and actual lake levels. The Mean Absolute Error (MAE) was **0.8000**, reinforcing that on average, the predicted values were within less than one unit of the actual values. However, the R-squared (R²) value was **-0.0082**, suggesting that the model did not explain the variance in the target variable better than a simple mean-based prediction. Despite low RMSE and MAE, the negative R² highlights limitations in the model’s explanatory power, which could be attributed to seasonality, external factors, or noise in the dataset not captured by the ARIMA(1,0,1) configuration.

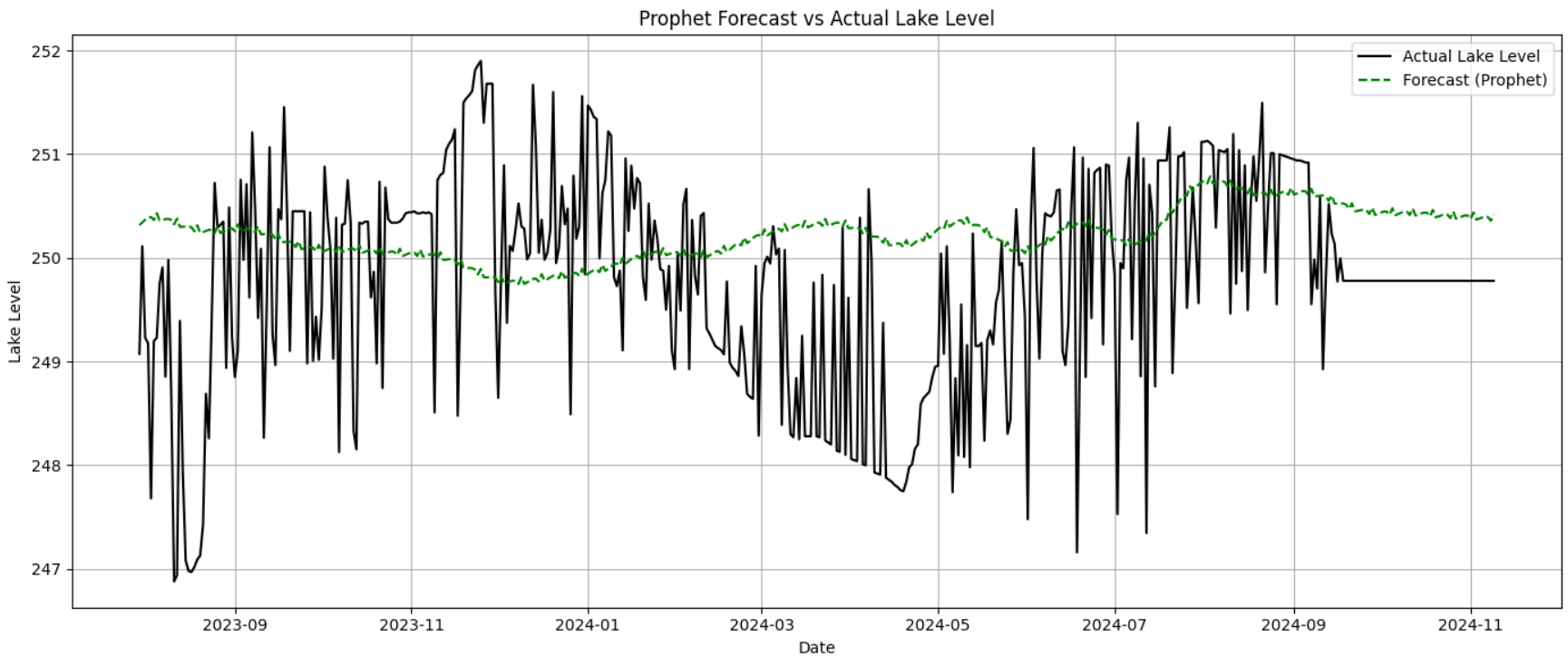


Figure 17: Prophet Model Fitting Check

|  |
| --- |
| Prophet Model Evaluation:  RMSE: 1.1590  MAE : 0.9058  R² : -0.2630 |

The Prophet model was rigorously evaluated to assess its forecasting accuracy for lake water levels. The dataset was split into training and testing sets using a 90/10 ratio, ensuring the last 10% of the time series was held out for evaluation. Prophet was trained on the historical data from 2012 onward and then tasked to predict lake levels for the test period. The resulting forecast was compared against the actual observed values using three standard performance metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²). The model achieved an RMSE of 1.1590, MAE of 0.9058, and an R² value of -0.2630. While the low RMSE and MAE suggest the predictions are close to actual values on average, the negative R² indicates that the model struggles to capture the variance in the data effectively compared to a naive mean predictor. This may be attributed to high-frequency variability or noise in the lake level data. A visual comparison further illustrates the performance, showing that while Prophet captures the overall direction and smooth trend of lake levels, it may underperform in modeling abrupt local fluctuations. Nonetheless, Prophet provides an interpretable and seasonality-aware forecast which remains a strong baseline for time series modeling.

# Conclusion

This project systematically analyzed and modeled time-series data related to lake water management, with a specific focus on forecasting lake levels and flow rates. Beginning with thorough data exploration, we assessed statistical summaries, identified outliers, and examined correlations between critical hydrological parameters such as rainfall, temperature, flow rate, and lake level. To ensure data integrity, anomalies were detected and addressed using Z-score and IQR methods, and stationarity was evaluated using the Augmented Dickey-Fuller (ADF) test. Autocorrelation and partial autocorrelation analyses further guided the selection of suitable forecasting models. Two prominent time-series models—**ARIMA** and **Facebook Prophet**—were implemented for forecasting both lake levels and flow rates. While ARIMA performed reasonably well in modeling historical trends, Prophet offered interpretability by decomposing trend and seasonality components. However, performance metrics such as RMSE, MAE, and R² indicated that while short-term predictions were moderately accurate, both models showed limitations in long-term generalization, particularly for the Prophet model in flow rate prediction. Overall, this study highlights the importance of comprehensive data preprocessing and model evaluation when working with environmental time-series data. Future work may benefit from integrating external variables (e.g., dam operations, regional climate patterns) or applying hybrid and deep learning models to improve forecasting accuracy and decision-making in smart water resource management systems.

# Future Work

While this project demonstrated the feasibility of using ARIMA and Prophet models for forecasting lake levels and flow rates, several areas remain open for future development. Incorporating **exogenous variables** such as dam operations, evaporation rates, and soil moisture content could enhance model accuracy and realism. **Advanced deep learning architectures** like LSTM or hybrid models that combine statistical and neural approaches may yield improved predictive performance, particularly for long-term horizons. Additionally, integrating **real-time sensor data** and building **interactive dashboards** using platforms like Streamlit or Power BI can provide dynamic decision support for water resource managers. Finally, extending the analysis to include **climate change scenarios** or **geospatial modeling** would help anticipate long-term impacts on lake ecosystems and inform sustainable water management strategies.